2016

Approaching Questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry

Vanessa Gorman
University of Nebraska-Lincoln, vgorman1@unl.edu

Robert J. Gorman
University of Nebraska-Lincoln, rgorman1@unl.edu

Research Article **Open Access**

Vanessa B. Gorman*, Robert J. Gorman

Approaching Questions of Text Reuse in Ancient Greek Using Computational Syntactic Stylometry

DOI 10.1515/opli-2016-0026
Received February 29, 2016; accepted October 14, 2016

Abstract: We are investigating methods by which data from dependency syntax treebanks of ancient Greek can be applied to questions of authorship in ancient Greek historiography. From the Ancient Greek Dependency Treebank were constructed syntax words (sWords) by tracing the shortest path from each leaf node to the root for each sentence tree. This paper presents the results of a preliminary test of the usefulness of the sWord as a stylometric discriminator. The sWord data was subjected to clustering analysis. The resultant groupings were in accord with traditional classifications. The use of sWords also allows a more fine-grained heuristic exploration of difficult questions of text reuse. A comparison of relative frequencies of sWords in the directly transmitted Polybius book 1 and the excerpted books 9–10 indicate that the measurements of the two texts are generally very close, but when frequencies do vary, the differences are surprisingly large. These differences reveal that a certain syntactic simplification is a salient characteristic of Polybius' excerptor, who leaves conspicuous syntactic indicators of his modifications.

Keywords: dependency syntax treebanking, text reuse, Greek historiography

1 Introduction

We are investigating methods by which data from dependency syntax treebanks of ancient Greek can be applied to questions of text reuse involving prose authors, with the ultimate goal of using computationally-developed evidence to evaluate text reuse cases important for the field of Greek historiography, including cases of epitomizing and excerpting. Our object is to both identify sources and estimate the accuracy of text reuse when writings by the source author are no longer extant in a direct transmission. Elsewhere (Gorman and Gorman 2014) we have developed a method to determine authorship based on micro-analyses of semantics and diction. This article represents a preliminary step towards the same goal, but using macro-analyses of syntactic data.

In this study we examine the relative frequency distributions of a certain type of pattern or sequence based on dependency syntax. Our goal is to establish whether such patterns provide a viable basis on which to identify associations computationally among objectively similar texts (e.g., texts by the same author or in the same genre). Thus, our study is a species of authorship attribution or verification. After a

Article note: This paper belongs to the special issue on Treebanking and Ancient Languages, ed. by Giuseppe G.A. Celano and Gregory Crane

***Corresponding author: Vanessa B. Gorman,** Dept. of History, University of Nebraska-Lincoln, Lincoln, NE United States, E-mail: vgorman1@unl.edu
Robert J. Gorman, Dept. of Classics and Religious Studies, University of Nebraska-Lincoln

discussion of the background literature in section 2, we begin section 3 by explaining the data structures we have constructed on the basis of dependency treebanks as the input for our attempts at attribution. Next, in section 4 we show the results of the preliminary test of our approach using the techniques of cluster analysis. Finally, in the discussion in section 4 we detail a few of the more significant syntactic features that come to light in the course of our investigation, in order to illustrate that, in contrast to, for example, character n-grams, the syntactic material that we propose has the advantage of being able to characterize stylistic distinctions in terms already familiar to linguists and Classicists. Section 5 concludes the paper by indicating that computational syntactic stylometry will be a fertile field of study in Greek prose generally and in historiography particularly.

2 Background Literature

Our investigation assumes the existence of what has been called a *stylome*, that is, "a set of measureable traits" of texts that "is extensive enough to distinguish between pairs of language users on the basis of their language use" (van Halteren et al. 2005: 66). In other words, individual language users—and especially authors of the highly elaborated literary works that are our focus—are assumed to have, as it were, stylistic fingerprints and, like physical fingerprints, these characteristics of style are unique and abiding. They constitute an "author invariant", a distinguishing mark by which authorship can be differentiated and identified. The search for such characteristics has a long history and, as one might expect, scholars have proposed many different candidates for the basis of analysis. Among the earliest suggested criteria were average word length or sentence length (Mendenhall 1887; further references at Grieve 2007: 252, sections 2.2.1 and 2.2.2). As computers became available to aid investigation, more extensive and complex data began to be examined for suitability (for a recent discussion with full bibliography, see Oakes 2014).

At the center of most attempts to find the stylistic fingerprint has been consideration of the author's vocabulary. For example, various measurements of vocabulary richness have been proposed. These measurements can be simple in concept, as in the Type-Token Ratio (the number of different words divided by total words in a text) or the ratio to the total of words appearing only once (*hapax legomena*) or twice (*dis legomena*) in a text. Although widely used, calculating the quality of an author's vocabulary involves non-trivial complications. In particular, the Type-Token Ratio and related measurements are affected strongly by the total length of the text. The frequency of new words decreases as total words increase, but the rate of decrease is difficult to account for. Thus, the literature is replete with formulae for capturing this phenomenon: Grieve (2007: 252–253) lists eleven measurements with reference to attribution studies that use the calculation of vocabulary richness in their work.

The value of these measurements of lexical richness—and of this stylistic feature more generally—remains subject to serious doubts. Tweedie and Baayen (1998: 324) present a detailed examination of the mathematical constants developed to measure vocabulary richness. Their study concludes that lexical constants should not be used to compare texts of different lengths. Hoover (2004: 152) reaches an even more negative conclusion, finding that measuring vocabulary richness is "of only marginal value" for authorship attribution.

These limitations have given impetus to the search for other *discriminators* by which authors can be distinguished. Oakes (2009: 1071), from whom we borrow this term, notes that the "overwhelming majority" of relevant studies rely on assessments of vocabulary richness and related lexical discriminators, and he specifies generally recognized characteristics integral to acceptable discriminators of all types. They should be of high frequency and good dispersion throughout a corpus, so that they can be effectively measured and compared. At the same time, they should be independent of content and not contingent on a work's genre or topic, for example. The best discriminators should not be subject to the conscious control of the author, but rather be subconscious choices. These criteria underlie several evaluations of distinctive stylometric features in the literature (Oakes 2009; Stamatatos 2009). Kestemont et al. (2016: 87) in particular emphasize the importance of psycholinguistic studies that identify language features that do not "actively attract cognitive attention" and thus can more objectively be judged as unconscious.

Three categories of language features are prominent among the discriminators that have been proposed to meet these criteria and overcome the weaknesses of vocabulary richness. In their famous study of the authorship of the *Federalist Papers*, Mosteller and Wallace (1963) demonstrated the value of authorship attribution based on measuring the frequency of function words. Roughly speaking, function words include prepositions, articles, conjunctions, auxiliary verbs, and other such classes that contain little referential meaning—they mean little in isolation—but, rather, represent grammatical relationships. Function words make up a closed set, since it is very difficult to create new members. They are to be contrasted with content words, such as nouns, verbs, and adjectives that carry clear referential or categorical meaning and, thus, vary significantly according to topic. They also constitute an open set, with neologisms appearing constantly. In contrast, it is easy to see how function words could be effective discriminators, since they are ubiquitous in all genres and less sensitive than content words to choice of topic. Chung and Pennebaker (2007: 347) point out that the average English speaker has a vocabulary of over 100,000 words, of which fewer than 400 are function words. But as for frequency, less than 0.04% of our vocabulary "accounts for over half the words we use". In addition, studies have shown that readers are not always conscious of the function words in a text (Drewnowski and Healy 1977; Schindler 1978). In general conversation, "we have virtually no control or memory over how and when they [i.e. function words] are used either by the speaker or ourselves" (Chung and Pennebaker 2007: 347). Because of their apparent advantages, function words have been used in many attribution studies in the years since Mosteller and Wallace (Baayen et al. 2002; Binongo 2003; Burrows 1987; Koppel et al. 2005).

More recently another kind of discriminator, implausible on the face of it, has taken center stage. Kjell (1994) turned from an analysis based on word units and introduced the surprising effectiveness of the character n-gram. This approach analyzes a text's letters or characters, including punctuation, numerals, and white space. These elements are usually measured as they appear in groups of multiple characters, called *digrams*, *trigrams*, etc. For example, the trigram of that word includes "tri", "rig", "igr", "gra", and "ram". A number of such analyses are represented in the literature (Houvardas and Stamatatos 2006; Peng et al. 2003; Koppel 2009; Stomatatos 2009). Studies based on sub-word n-grams can often achieve good success. Juola (2006: 296–297) reports the results of an "*Ad-hoc* Authorship Attribution Competition" sponsored by learned societies in the field: "methods based on simple lexical statistics tended to perform substantially worse than methods based on N-grams".

Syntactic features seem easily to meet the criteria of sufficient frequency and dispersion, as well as unconscious use (Stamatatos 2009). They have been studied with some success, beginning with Baayen et al. (1996). The analysis works from a syntactically annotated corpus; so that the same techniques generally applied to word frequencies might also be used for syntax, "pseudo-words" were constructed based on re-write rules applied to the underlying data (Baayen et al. 1996: 123–124). The authors indicate the success of their approach, noting that, while function words provide an economical way to take syntax into account, more direct measurements of syntactic features lead to "higher discriminatory resolution" (Baayen et al. 1996: 129). In spite of this potential, attribution studies based on units of syntax are relatively infrequent (see the literature review in Stamatatos 2009). This rarity is due to some degree to questions of economy. Syntactic analyses of this type require an annotated corpus and a preparatory step of some complexity. Besides the time and expense entailed, automated annotation or parsing can introduce significant error that can degrade the results of attribution analysis.

3 Method

In spite of the difficulties sometimes associated with this approach, we have chosen to use syntactic features as the discriminator upon which we base the present investigation. In recent years, a substantial body of syntactically annotated ancient Greek texts has become freely available. These texts, provided by the Ancient Greek Dependency Treebank (AGDT), under the auspices of the Perseus Digital Library and Leipzig University, are hand annotated and can be considered reliably accurate and consistent. The use of this corpus for our data may allow us to avoid the shortcomings noted in the scholarship.

Before turning to an explication of our method and results, it is necessary to review the annotation scheme of the AGDT and illustrate the grammatical information recorded there. A look at the annotation for a simple sentence from Book 2 of the historian Polybius will serve our purpose: ἐποίει δὲ τοῦτο διὰ τὴν στενότητα τῶν τόπων ('They were doing this thing because of the narrowness of the locations', Polyb. 2.66.10). An dependency analysis using the AGDT tag set 1.0 produces the dependency syntax tree, found in figure 1:

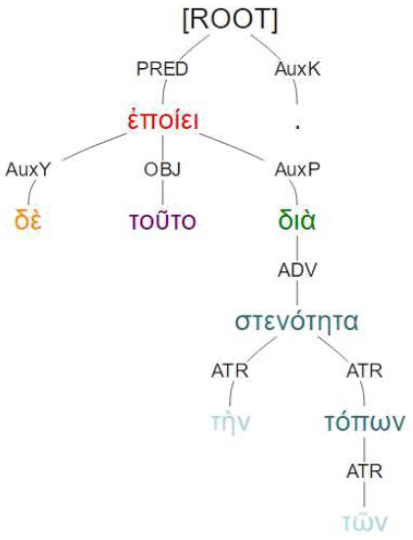

Figure 1.

When viewed in XML, it looks like this:

```
<sentence document_id="urn:cts:greekLit:tlg0543.tlg001.perseus-grc1" subdoc="2.66.9">
    <word id="1" form="ἐποίει" lemma="ποιέω" postag="v3siia---" relation="PRED" head="0"/>
    <word id="2" form="δὲ" lemma="δέ" postag="d--------" relation="AuxY" head="1"/>
    <word id="3" form="τοῦτο" lemma="οὗτος" postag="p-s---na-" relation="OBJ" head="1"/>
    <word id="4" form="διὰ" lemma="διά" postag="r--------" relation="AuxP" head="1"/>
    <word id="5" form="τὴν" lemma="ὁ" postag="l-s---fa-" relation="ATR" head="6"/>
    <word id="6" form="στενότητα" lemma="στενότης" postag="n-s---fa-" relation="ADV" head="4"/>
    <word id="7" form="τῶν" lemma="ὁ" postag="l-p---mg-" relation="ATR" head="8"/>
    <word id="8" form="τόπων" lemma="τόπος" postag="n-p---mg-" relation="ATR" head="6"/>
    <word id="9" form="." lemma="punc1" postag="u--------" relation="AuxK" head="0"/>
</sentence>
```

For each lexical token, the following data points are available: 1) position in linear order; 2) word form as it appears in the text; 3) lemma; 4) morphological classification (i.e., the part-of-speech tag for token 3 is p-s---na-, indicating that τοῦτο is a singular neuter accusative pronoun); 5) syntactic role (i.e., the AuxY of token 2 signifies sentence adverbial); and 6) dependency parent. Thus, there is a wealth of information with which to fuel our investigation.

Once the decision was taken to use syntactic data, there remained a choice among many possible procedures. Van Halteren et al. (2005: 71) describe an ideal set of syntactic discriminators comprehending "a rich notion of syntax, which includes hierarchical relations, co-reference, long-distance dependencies, etc.". Unfortunately, such an annotated data set is not currently feasible for most corpora. Syntactic-based authorship analyses thus fall back on simpler features. As we have seen, Baayen et al. (1996) use syntactic

re-write rules. Gamon (2004) opts for a similar approach supplemented with part-of-speech trigrams. Luyckx and Daelemans (2005) rely on a process of "shallow text analysis" that identifies parts-of-speech, noun, verb, and prepositional phrases, and subjects and objects of verbs. Uzuner and Katz (2005) analyze features including sentence-initial or sentence-final structures used to shift focus, as well as an estimate of sentence complexity calculated from the depth of subordination of phrases and clauses. For our part, we have decided to base our study on one of the most salient features of dependency analysis, the series of one-to-one subordinate relationships linking each word in a sentence with the sentence root. On this basis we have constructed a discriminator for each word in our corpus and have for convenience named this construction the *syntax word* or *sWord*.

Following suggestions in the literature,[1] we constructed syntax words by tracing the shortest (or geodesic) path from each leaf node to the root for each sentence tree. These paths supply the sequences according to which the constituent elements of the sWords are arranged. The elements themselves are chosen from the syntactical annotations available for each token in the treebank, the most important being dependency relationship and morphological classification. Thus, for example, the syntax words corresponding to the tokens in our Polybian example sentence might look like this:

ἐποίει: <sWord id="1"> **#-PRED** </sWord>
δὲ: <sWord id="2"> **#-PRED-AuxY** </sWord>
τοῦτο: <sWord id="3"> **#-PRED-OBJ** </sWord>
διὰ: <sWord id="4"> **#-PRED-AuxP** </sWord>
τὴν: <sWord id="5"> **#-PRED-AuxP-ADV-ATR** </sWord>
στενότητα: <sWord id="6"> **#-PRED-AuxP-ADV** </sWord>
τῶν: <sWord id="7"> **#-PRED-AuxP-ADV-ATR-ATR** </sWord>
τόπων: <sWord id="8"> **#-PRED-AuxP-ADV-ATR** </sWord>

To interpret: the sWord of the fifth token, τὴν, represents it as an attribute (ATR) specifying an adverbial (ADV), that is introduced by a prepositional bridge structure (AuxP), that ties it to the main verb (PRED), that depends on (or is equivalent to) the hypothetical root (#) of the sentence. It is important to note that the sWord provides a "backbone structure" that can be expanded through the addition of whatever information is available in the annotation. In particular, in the following analysis, we have supplemented the dependency relationship with the part-of-speech tag for each word. The tested version of the first words of our sample text is thus:

ἐποίει: <sWord id="1"> **#-PRED-v** </sWord>
δὲ: <sWord id="2"> **#-PRED-v-AuxY-d** </sWord>
τοῦτο: <sWord id="3"> **#-PRED-v-OBJ-p** </sWord>

Here, instead of simple **#-PRED** for the initial word and sentence root, the part-of-speech is indicated by the additional tag **-v** (for verb). Word 2 is additionally identified as an adverb (**-d**), and word 3 as a pronoun (**-p**).

Computationally, the syntax word has the same advantages that led Baayen et al. (1996) to create pseudo-words based on syntactic re-write rules: they can be produced from the underlying data in a relatively straightforward way, and they can serve as input to the wide variety of analytic procedures currently applied to normal lexical words. In addition, from our point of view an important advantage of the syntax word is

1 Examples in the literature of dependency syntax used for authorship attribution and verification are quite rare. For a recent and detailed review of the relevant material, see Hollingsworth (2012a: 3–6). The scarcity of syntactic approaches is not surprising given the necessity for syntactically annotated data as input. In a survey of computational studies of authorship, Stamatos (2009: 542–543) discusses the drawbacks of reliance on relatively faulty taggers and parsers. Our study has the advantage of a sufficiently large corpus of ancient Greek texts that have been annotated by hand. We have been inspired to test the usefulness of sWords for our purposes by the work of Hollingsworth (2012a and 2012b) and Sidorov et al. (2012). These scholars base their studies on units composed of sequences of syntactic relationships arranged hierarchically according to position on the geodesic in a dependency tree.

its relative transparency to classicists who lack training in computational analysis. In general, dependency syntax—emphasizing as it does the relationship between pairs of individual words, rather than between words and intermediate constituents—is close to the traditional grammatical analyses common in classical studies. Syntax words built from dependency trees should thus have considerable heuristic value within the field, since they encode familiar relationships (dependencies) in a straightforward way and therefore allow stylistic observations based on them to be recast in more traditional terms. We expect that, when we can explain and illustrate the characteristic patterns of sWords for a given author using the familiar categories of philological research, computational results are more likely to have an impact on the discipline.

Using as input sWord data derived from texts associated with the AGDT, we performed cluster analyses in order to determine whether sWords would constitute a suitable basis for effective computational analysis. In this case, the sWords were constructed from the syntactical relationship and part-of-speech data for each token along the geodesic from root to target word. Using the R statistical programming language, we produced a table giving the relative frequency of each variable for each text. The relative frequency table includes the 2504 most common sWords in the corpus of 36 files. The decision to consider 2504 of the total 93,867 variables is arbitrary: the selected variables represent all sWords with a corpus-wide mean relative frequency of at least 0.0025. The number of unique sWords in the corpus was 93,867; the total token count was 520,889. From the relative frequency table is then generated a dissimilarity matrix giving, pairwise, a measure of distance between every two texts in the corpus. This distance measure may be calculated in several ways. We chose Euclidean distance, for which the formula is:

$$d(p,q) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 \ldots + (q_n - p_n)^2}$$

Using this calculation, the large number of variables examined in our corpus are reduced to a single figure for each pair of texts.[2]

These numbers then become input in an algorithm for hierarchical clustering. This procedure uses mathematical formulae to arrange objects into "clusters" of similar items.[3] The results are easiest to read in a dendrogram. The clustering of our data is reproduced in figure 2.

4 Results and Discussion

The results of the clustering analyses were generally successful. For example, the major branches between which the algorithm finds the greatest difference represent the distinction between prose and verse. A second main branch divides epic from the other poetic genres. More generally, according to the dissimilarity matrix (Figure 3), the average difference among all prose files is 59.45, among verse files 67.83, and between the prose and verse groups 76.27. Interestingly, the most disparate prose files are Diodorus Siculus Book 11 and the selections from the orations of Lysias (72.44). The most syntactically distinct verse files according to this measure are the *Trachiniae* of Sophocles and Hesiodic *Shield of Heracles* (88.89). Among all files, the *Shield of Heracles* and Plutarch's *Life of Lycurgus* present the apogee of syntactic diversity (91.69).

At the other end of the spectrum, analysis based on sWords seems sufficiently fine-grained to correctly associate various parts of the *Iliad* and the *Odyssey*. These large works were each divided into six chunks of

2 Before calculating the dissimilarity scores, we have scaled, or standardized, the relative frequencies for each variable. This is done by, first, subtracting the mean frequency of a given variable from each observation of the variable and, second, dividing frequencies resulting from the first step by the standard deviation of the variable. The result is that each variable has a mean value of 0 and a standard deviation of 1. The purpose of standardization is compensate for the large difference in value among variables. For example, the mean relative frequency of the sWord **#-PRED-v** (the main verb of a non-coordinated sentence) is 4.504. In contrast, the mean of the frequency of the 135th most common sWord, **#-PRED-v-ADV-v-OBJ-p** (a prepositional object of a verb adverbially modifying the main verb), is 0.065. The two numbers differ by nearly two orders of magnitude. Correspondingly, Euclidean distances measured for the first variable are likely to be much larger than those for the second variable, with the potential for exaggerated influence on the total distance calculation.

3 Our analysis used R 3.2.2 (R Core Team 2015). Data were extracted from our corpus of sWords with the R package XML (Lang et al. 2015). The clustering algorithm was "hclust()" with the method parameter "ward.D2". For a discussion of the mathematical basis of the algorithm, see Murtagh and Legendre (2014).

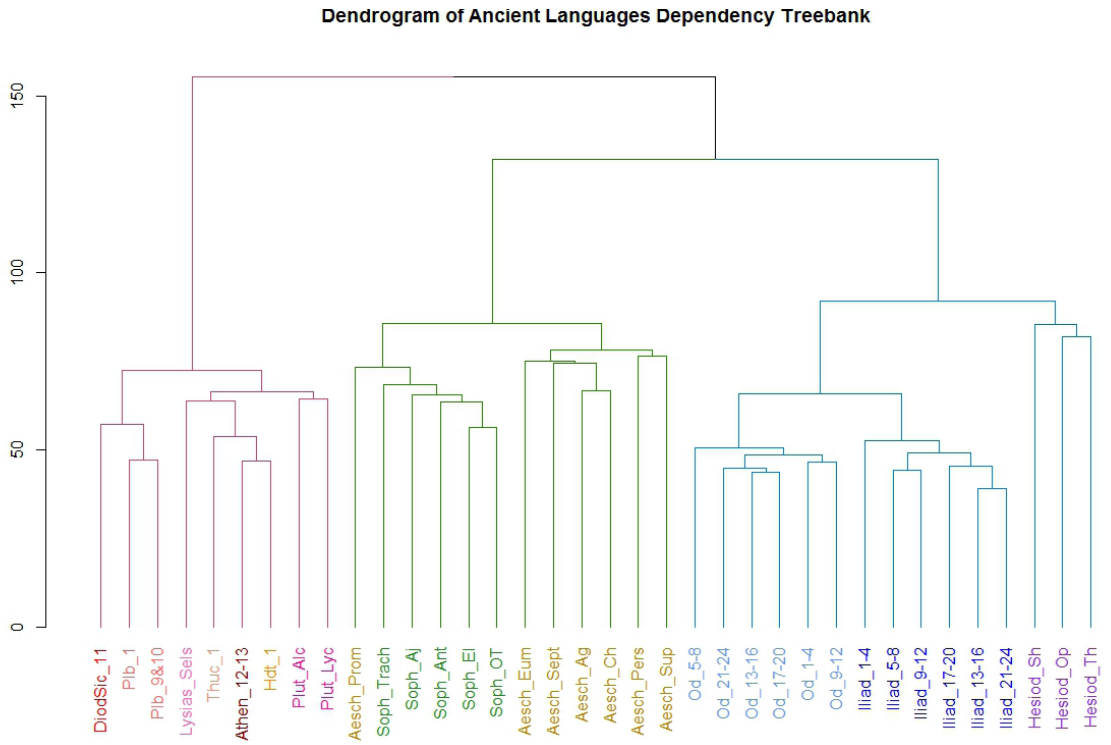

Figure 2.
Abbreviations (in order from top of graph): DiodSic_11 = Diodorus Siculus, *Bibliotheca Historica* Book 11; Plb_1, Plb_9&10 = Polybius, *Histories* Book 1, Books 9 and 10; Lysias_Sels = Lysias, *Orations* 1, 12, 14, 15, 23; Thuc_1 = Thucydides, *Histories* Book 1; Athen_12–13 = Athenaeus, *Deipnosophistae* Books 12–13; Hdt_1 = Herodotus, *Histories* Book 1; Plut_Alc, Plut_Lyc = Plutarch, *Life of Alcibides, Life of Lycurgus*; Aesch_Prom = Aeschylus, *Prometheus Bound;* Soph_Trach, Soph_Aj, Soph_Ant, Soph_El, Soph_OT = Sophocles, *Women of Trachis, Ajax, Antigone, Electra, Oedipus the King;* Aesch_Eum, Aesch_Sept, Aesch_Ag, Aesch_Ch, Aesch_Pers, Aesch_Sup = Aeschylus, *Eumenides, Seven Against Thebes, Agamemnon, Libation Bearers, Persians, Suppliants;* Od_5–8 etc. = Homer, *Odyssey* (books as given); Iliad_1–4 etc. = Homer, *Iliad* (books as given); Hesiod_Sh, Hesiod_Op, Hesiod_Th = Hesiod, *Shield of Heracles, Works and Days, Theogony.* Prose works are labelled in reddish hues, dramatic poetry in greens, and epic poetry in blues.

four books. The clustering algorithm correctly recognizes as different these segments and separates them into the appropriate groupings. The average distance among chunks of the *Iliad* is 46.26 (min. = 38.99, max. = 51.01), while within the *Odyssey* it is 46.88 (min. = 43.81, max. = 50.12). Between the two works, the relevant measures are average = 50.31, minimum = 46.2, and maximum = 54.32 (*Iliad* 17–20 v. *Odyssey* 9–12).

Of particular interest is the clustering of Aeschylus' *Prometheus Bound* ("Aesch_Prom" in the figures). This work has been grouped not with the other plays of Aeschylus, but among the works of Sophocles. This arrangement might at first glance seem to be a weakness in the clustering results, but in fact experts in Greek tragedy have long doubted the authenticity of the *Prometheus Bound* as a work of Aeschylus (Griffith 1977). The treatment of the *Prometheus* by the clustering algorithm thus provides independent support for that doubt while serving as further evidence for the sensitivity of the sWord as a stylometric discriminator.

Not surprisingly, the clustering experiment also reveals certain limits in this approach to syntactic analysis. We see these limitations in the grouping of the two files associated with the historian Polybius ("Plb_1" and "Plb_9&10" in the figures). To understand the problem, some background information is necessary: writing in the 2nd century BCE, he was addressing a Greek audience in order to explain the growth of Roman power during the Punic Wars. Of the 40 books that comprised his *Histories*, only books 1–5 have come to us through a direct manuscript transmission, while the others arrive in larger or smaller pieces, mostly through Byzantine excerptors and epitomizers, working a millennium or more later. For our study, we chose to compare the syntactic analysis of book 1 to that of books 9–10, that derive from a much later compilation, the *Excerpta Antiqua*, dating from sometime before 1000 CE (Moore 2011; Pearse 2013).

Distance Matrix

	Aesch_Ag	Aesch_Ch	Aesch_Eum	Aesch_Pers	Aesch_Prom	Aesch_Sept	Aesch_Sup	Athen_12-13	DiodSic_11	Hdt_1	Hesiod_Op	Hesiod_Sh	Hesiod_Th	Iliad_1-4	Iliad_13-16	Iliad_17-20	Iliad_21-24	Iliad_5-8	Iliad_9-12	Lysias_Sels	Od_1-4	Od_13-16	Od_17-20	Od_21-24	Od_5-8	Od_9-12	Plb_1	Plb_9&10	Plut_Alc	Plut_Lyc	Soph_Aj	Soph_Ant	Soph_El	Soph_OT	Soph_Trach
Aesch_Ag																																		
Aesch_Ch	66.83																																	
Aesch_Eum	70.73	74.96																																
Aesch_Pers	72.96	76.65	78.31																															
Aesch_Prom	69.92	74.18	75.13	76.98																														
Aesch_Sept	70.00	75.09	75.13	76.36	74.59																													
Aesch_Sup	71.19	74.27	77.93	76.55	73.39	76.77																												
Athen_12-13	67.06	70.66	69.51	75.07	70.34	73.43	72.08																											
DiodSic_11	84.95	86.78	86.52	87.87	87.19	87.68	90.09	64.70																										
Hdt_1	72.85	74.24	74.26	76.07	74.35	75.30	77.92	47.02	58.24																									
Hesiod_Op	79.02	83.50	83.69	83.56	82.04	83.83	85.53	74.90	87.61	75.99																								
Hesiod_Sh	84.53	86.76	87.87	87.97	87.02	86.94	88.75	81.55	91.45	82.41	87.02																							
Hesiod_Th	80.55	83.29	84.30	81.49	81.62	81.36	84.21	76.74	87.71	76.96	81.85	82.04																						
Iliad_1-4	71.54	73.86	74.20	75.72	72.41	73.32	76.80	64.52	79.35	64.51	73.52	76.50	72.02																					
Iliad_13-16	67.84	69.45	68.70	71.71	68.46	69.32	72.78	59.49	73.01	58.74	69.24	72.19	67.82	48.70																				
Iliad_17-20	68.55	71.63	72.82	74.54	70.69	72.79	73.60	61.95	76.72	62.28	71.76	72.30	70.08	51.02	43.85																			
Iliad_21-24	66.85	69.63	70.97	71.88	68.64	70.09	72.79	60.72	73.78	59.43	68.10	71.35	67.51	49.21	39.00	44.14																		
Iliad_5-8	68.93	72.11	72.44	74.35	70.85	71.53	73.68	62.03	75.38	60.39	71.95	74.19	70.26	49.48	43.65	48.23	45.11																	
Iliad_9-12	69.76	71.88	72.04	74.31	70.05	71.73	73.86	62.33	76.32	62.27	71.57	75.05	69.91	49.68	43.83	48.04	45.67	44.40																
Lysias_Sels	76.80	79.28	80.42	83.94	79.29	82.10	83.19	60.22	72.44	59.57	83.30	89.31	83.67	71.12	68.12	70.00	67.88	69.43	69.63															
Od_1-4	68.36	71.61	73.46	72.26	69.47	71.08	73.66	60.38	76.53	59.94	71.45	76.02	69.76	53.67	50.08	53.48	50.11	49.45	51.41	67.26														
Od_13-16	68.92	69.97	71.92	71.94	69.70	69.90	73.99	61.31	74.89	60.09	70.44	76.33	69.60	50.87	47.09	52.49	47.81	46.21	48.41	68.00	44.07													
Od_17-20	66.48	69.73	71.11	71.52	68.35	70.89	72.44	59.26	75.79	59.96	67.45	75.44	69.90	51.61	48.21	50.86	47.20	47.73	48.91	66.59	46.49	43.82												
Od_21-24	70.05	72.73	73.07	73.21	70.68	71.27	74.42	62.99	77.10	61.53	71.85	77.51	70.65	53.57	49.15	54.01	48.89	49.60	49.94	69.93	48.06	43.97	45.42											
Od_5-8	70.74	72.72	74.25	74.15	71.68	72.93	74.60	63.39	78.42	64.03	71.88	76.74	71.01	54.09	49.71	50.43	49.65	49.14	50.58	70.56	50.04	46.42	48.00	50.13										
Od_9-12	71.00	73.43	73.70	74.64	71.88	72.52	76.28	64.46	77.97	63.39	72.71	76.60	69.84	53.22	50.19	54.32	49.78	48.57	50.86	70.77	46.47	44.66	48.41	48.23	49.16									
Plb_1	77.07	80.37	79.31	82.24	80.24	81.39	82.06	56.26	51.56	56.00	81.49	86.79	81.79	72.67	67.47	71.26	68.67	70.66	71.09	65.81	70.55	69.19	69.86	70.95	72.37	72.33								
Plb_9&10	77.12	80.67	79.94	83.31	81.17	81.63	82.74	56.57	58.01	55.33	82.72	88.64	84.32	73.28	68.71	71.05	68.93	70.63	71.87	63.01	70.98	69.11	69.89	71.14	72.36	73.53	47.25							
Plut_Alc	82.50	83.20	83.84	86.25	84.31	85.62	87.37	62.57	62.66	58.35	83.74	91.02	86.30	75.64	70.56	73.51	70.74	72.87	73.52	68.10	72.11	71.60	71.06	73.45	74.17	74.02	61.03	61.49						
Plut_Lyc	79.67	82.68	81.58	85.96	83.70	84.50	86.17	58.65	70.08	59.96	86.02	91.70	86.56	77.10	72.84	75.01	73.92	74.86	75.90	67.27	73.80	73.58	73.66	75.76	76.28	76.03	61.46	60.42	64.37					
Soph_Aj	66.54	69.68	73.32	75.69	69.86	73.28	73.95	65.35	82.22	68.63	77.64	83.97	79.72	67.44	62.66	65.21	63.01	66.16	64.78	72.12	65.52	64.45	62.86	65.83	67.27	67.16	74.72	74.71	78.43	78.10				
Soph_Ant	68.00	73.00	73.21	72.61	69.88	74.31	75.40	65.34	82.00	67.87	77.83	85.47	81.25	68.48	66.49	67.91	65.14	67.07	66.20	72.80	66.99	65.42	63.58	67.33	67.93	69.30	75.73	75.71	78.25	77.76	65.27			
Soph_El	66.82	71.73	73.02	75.75	68.87	72.69	72.14	67.18	86.02	70.69	78.78	86.90	81.38	69.79	66.77	67.44	65.84	68.13	67.52	73.89	66.95	66.75	64.32	67.48	68.51	69.85	77.71	77.21	81.46	80.88	64.00	62.24		
Soph_OT	64.01	69.69	72.23	73.50	65.81	72.43	71.47	62.21	82.86	66.60	77.00	84.25	80.43	67.16	64.36	64.01	63.28	65.22	64.96	70.33	64.64	64.01	62.16	65.64	67.05	67.49	74.50	74.65	77.92	76.72	62.38	61.57	56.42	
Soph_Trach	73.46	75.27	77.52	77.79	73.59	76.65	76.42	70.82	86.92	72.30	82.18	88.99	84.17	72.09	68.72	71.22	68.47	70.48	70.83	77.34	69.03	68.15	66.22	69.34	70.57	72.29	79.95	79.48	82.68	83.33	68.79	68.48	63.39	64.05
Thuc_1	75.69	77.60	77.47	78.43	77.43	78.38	81.76	56.21	56.93	47.89	77.25	83.88	78.75	66.10	61.06	64.57	61.16	63.36	64.21	59.32	62.98	62.77	62.00	64.36	65.07	65.14	55.73	55.48	58.44	61.94	71.40	71.74	73.99	70.31

Figure 3

These books were selected arbitrarily among the indirectly transmitted works as having a comparable sample size (28,288 tokens in book 1 compared to 26,688 in books 9–10).

Ex hypothesi, one might expect a marked difference between the directly transmitted book 1 and the books which have passed through the hands of excerptors and compilers. However, the cluster analyses diagrammed in Figure 2 tells us that the two files in question (book 1 v. books 9 and 10) are more similar than any two other single-author prose files in the corpus (distance = 47.24) and the second most similar of all comparisons between two authors.[4] Nonetheless, while the sensitivity of the distance matrix and the clustering algorithm may not be fine-grained enough for the subtle differences involved in text reuse, the sWord data has the advantage of allowing significant distinctions to be revealed by manual inspection.

For this exploration, we rely upon a simple spreadsheet containing for each file in the corpus two measurements per sWord*:* relative frequency and z-score (see above, note 2). Here we use this calculation simply to make our data easier to understand and analyze by hand, since raw frequencies can be very difficult to interpret. Turning now to the data from the two Polybius files, it is apparent from the distance matrix and clustering dendrogram that the relative frequencies in Polybius 1 are generally very close to those in books 9 and 10. However, when frequencies do vary, the differences are surprisingly large for parts of a single work by the same author. For example, if we assemble the z-scores most divergent between Polybius 1 and Polybius 9–10, we find that the average difference for the top 25 scores is 1.397 standard deviations. To put this finding into perspective, this difference is greater than the one existing between the first books of Herodotus and Thucydides (1.104 standard deviations).

With this background in mind, we can isolate some of the notable syntactic differences between the two Polybius texts. Perhaps most striking is the sWord #-PRED-v-ADV-v, the seventh most frequent syntax word in the corpus. The z-score for this structure in Polybius 1 is 1.919 (427 occurrences in 26,894 tokens), but in Polybius 9–10 it is -0.626 (218 occurrences in 19,378 tokens), giving a z-score difference of more than 2.5 standard deviations. At this point, it is important to remember that sWords are agglutinative constructions, so that the possible number of the unit #-PRED-v-ADV-v depends on the number of parent structures #-PRED-v (this sWord represents the main verb of a sentence). Thus, to get a truer picture of how characteristic #-PRED-v-ADV-v is of the style of Polybius 1 against Polybius 9–10, we must divide the frequency of #-PRED-v-ADV-v by the frequency of #-PRED-v. The results confirm the distinguishing nature of this sWord. In Polybius 1, #-PRED-v-ADV-v is 66.3% as frequent as #-PRED-v (z-score = 3.444); in Polybius 9–10, the ratio is 42% (z-score = 1.107). We can conclude that while both texts show a preference for #-PRED-v-ADV-v in comparison to the corpus as a whole, for Polybius 1, where the frequency is unusually high, the sWord should be considered a distinctive feature.

We have claimed heuristic and pedagogical transparency for sWords. What then is the #-PRED-v-ADV-v when cast in traditional terms? As the abbreviations suggest, ADV-v indicates a verb (-v) depending directly on the main verb as an adverbial. This structure can be manifested in three ways. The first is as a circumstantial participle: Polybius 1.2.5, μετὰ δὲ ταῦτα προσέλαβον τὴν τῆς Ἀσίας ἀρχήν, καταλύσαντες τὴν τῶν Περσῶν δυναστείαν (‘Afterwards, they **seized** control of Asia, **having broken** the power of the Persians’). The second is often considered a variety of the first, namely the genitive absolute construction: Polybius 1.7.12, ὧν ἀναπεμφθέντων εἰς τὴν Ῥώμην, οἱ στρατηγοὶ ... ἅπαντας κατὰ τὸ παρ’ αὐτοῖς ἔθος ἐπελέκισαν .... (‘These men **having been sent** to Rome, the consuls ... **beheaded** them all in accordance with their custom’). Third is the adverbial use of a relative clause with no antecedent (or the antecedent is attracted into the relative: Polybius 1.7.6, Ῥηγῖνοι γάρ, καθ’ ὃν καιρὸν Πύρρος εἰς Ἰταλίαν ἐπεραιοῦτο ... ἐπεσπάσαντο φυλακὴν... (‘The Rhegians, on which occasion Pyrrhus **crossed** to Italy, ... **begged** for a garrison ...’). Examination of the dependency trees in which #-PRED-v-ADV-v occurs shows that it is the participial construction that makes up a large majority of the relevant examples (circumstantial participle = 77.3% of examples of #-PRED-v-ADV-v; genitive absolute = 21.3%; relative clause = 1.3%). In fact, the pattern identified here becomes even stronger when we include 1) single participles dependent on coordinated predicates and 2) coordinated

4 It is fascinating that the closest clustering among all prose files is between Herodotus Book 1 and Athenaeus Books 12–13 (distance = 47.02). The text of Athenaeus is a pastiche of earlier authors, including passages from Herodotus himself, so it is difficult to interpret the significance of this similarity without further investigation.

participles on single predicates. Each of these structures has a different sWord*:* participle on coordinated predicates = #-COORD-c-ADV-v; coordinated participles on single predicate = #-PRED-v-COORD-c-ADV_CO-v. Taken together, the ratio of circumstantial participles (including genitive absolutes) to all main verbs in Polybius 1 is more than 75%. The prominence of this construction as a defining feature of Polybian style is apparent in its z-score of 3.89.

Thus, we suggest that by using sWords in this way we can develop a quantified description of a text's syntactic peculiarities. This information can be brought to bear on research problems such as questions of text reuse accuracy. For example, based on this preliminary analysis of Polybius, we may conclude that the excerptor is generally quite accurate. However, he has reworked the original, reducing the features most sharply Polybian and farthest from a more standard prose style. This process of normalization has left conspicuous indicators of the excerptor's modifications. Closer considerations of the particular characteristics of the differences may ultimately lead us to a better understanding of the relationship between the texts and a greater ability to detect modification in reuse. These developments may in turn result in a more precise judgment about the accuracy of text reuse more generally.

5 Conclusion

The purpose of this paper is to present evidence for the feasibility of syntacto-stylistics using the syntax word. We have sketched the advantages of this approach from a purely computational standpoint, where the sWord may offer the basis for a method of authorship attribution that is robust against shared content, since it does not rely directly on vocabulary choice. The good results of the clustering experiment indicate that analysis using the syntax word is indeed viable and more work in this direction is warranted. At the same time, we have suggested that the sWord may be used to help illustrate the predilections of our texts in a more traditional way; perhaps we will someday have the analog of the lexical concordance to guide us to a detailed understanding of each author's particular syntax. In any event, the results of our work in this area—extremely tentative as they may be—continue to foster in us the belief that treebanks and other annotated corpora will eventually become indispensable resources in the study of classical history and literature.

References

Baayen, Harald, Hans van Halteren & Fiona Tweedie. 1996. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing* 11(3). 121–131.

Baayen, Harald, Hans van Halteren, Anneke Neijt & Fiona Tweedie. 2002. An experiment in authorship attribution. *Journées internationales d'Analyse statistique des Données Textuelles* 6. http://www.sfs.uni-tuebingen.de/~hbaayen/publications/BaayenVanHalterenNeijtTweedieJADT2002.pdf (accessed 1 August 2016).

Binongo, José G. 2003. Who wrote the 15th book of Oz? An application of multivariate analysis to authorship attribution. *Chance* 16(2). 9–17.

Burrows, J. F. 1987. Word-patterns and story-shapes: the statistical analysis of narrative style. *Literary and Linguistic Computing* 2(2). 61–70.

Chung, Cindy K., James W. Pennebaker. 2007. The psychological function of function words. In Klaus Fiedler (ed.), *Social communication*, 343–359. New York: Psychology Press

Drewnowski, Adam, Alice F. Healy. 1977. Detection error on the and and: evidence for reading units larger than the word. *Memory and Cognition* 5(6). 636–647.

Eder, Maciej. 2015. Does size matter? Authorship attribution, short samples, big problem. *Digital Scholarship in the Humanities* 30. 67–182

Gamon, Michael. 2004. Linguistic correlates of style: authorship classification with deep linguistic analysis features. In Lothar Lemnitzer, Detmar Meurers & Erhard Hinrichs (eds.), *Proceedings of the 20th international conference on computational linguistics*, 611–617. Stroudsburg, PA: Association for computational linguistics. http://delivery.acm.org/10.1145/1230000/1220443/p611-gamon.pdf?ip=129.93.224.1&id=1220443&acc=OPEN&key=B63ACEF81C6334F5%2EEE2BA0AAC6332229%2E4D4702B0C3E38B35%2E6D218144511F3437&CFID=650791781&CFTOKEN=19509558&__acm__=1470060609_94d9b24140627b52e6626c0f8fee2bfd (accessed 1 August 2016),

Gorman, Robert J., Vanessa B. Gorman. 2014. *Corrupting luxury in ancient Greek literature.* Ann Arbor, MI: Michigan University Press.

Grieve, Jack. 2007. Quantitative authorship attribution: an evaluation of techniques. *Literary and Linguistic Computing* 22(3). 251–270.

Griffith, Mark. 1977. *The Authenticity of 'Prometheus Bound'.* Cambridge: Cambridge University Press.

Hollingsworth, Charles. 2012a. Syntactic stylometry: using sentence structure for authorship attribution. MA thesis. Athens, Georgia: University of Georgia. http://www.ai.uga.edu/sites/default/files/theses/hollingsworth_charles_d_201208_ms.pdf (accessed 16 June 2016).

Hollingsworth, Charles. 2012b. Using dependency-based annotations for authorship identification. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds), *Text, speech and dialogue*, 314–319. Berlin: Springer. http://link.springer.com/chapter/10.1007/978-3-642-32790-2_38 (accessed 28 February 2016).

Hoover, David L. 2003. Another perspective on vocabulary richness. *Computers and the Humanities* 37(2). 151–178.

Houvardas, John, Efstathios Stamatatos. 2006. N-gram feature selection for authorship identification. In Jérôme Euzenat and John Domingue (eds), *Artificial intelligence: methodologies, systems, and applications* (*AIMSA* 2006), 77–86. Berlin: Springer.

Juola, Patrick. 2006. Authorship attribution. F*oundations and Trends in Information Retrieval* 1(3). 233–334.

Kestemont, Mike. 2014. Function words in authorship attribution: From black magic to theory? In *Proceedings of the 3rd workshop on computational linguistics for literature* (*CLfL*), 59–66. Stroudsburg, PA: Association for computational linguistics.

Kestemont, Mike, Justin Stover, Moshe Koppel, Folgert Karsdorp & Walter Daelemans. 2016. Authenticating the writings of Julius Caesar. *Expert Systems with Applications* 63. 86–96.

Kjell, B. 1994. Discrimination of authorship using visualization. *Information Processing and Management* 30(1). 141–15.

Koppel, Moshe, Jonathon Schler & Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data* mining, 624–628. New York: Association for computing machinery. http://dl.acm.org/citation.cfm?id=1081947 (accessed 1 August 2016).

Koppel, Moshe, Jonathon Schler & Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology* 60(1). 9–26.

Lang, Duncan Temple, and the CRAN Team. 2015. *XML: Tools for parsing and generating XML within R and S-Plus*. R package version 3.98–1.3. http://CRAN.R-project.org/package=XML (accessed 28 February 2016).

Luyckx, Kim, Walter Daelemans. 2005. Shallow text analysis and machine learning for authorship attribution. In T. van der Wouden (ed.), *Proceedings of the fifteenth meeting of computational linguistics in the Netherlands*, 149–160. http://www.clips.ua.ac.be/bibliography/shallow-text-analysis-and-machine-learning-for-authorship-attribution (accessed 1 August 2016).

Mendenhall, T. C. 1887. The characteristic curves of composition. *Science* 11. 237–249.

Meyer, David, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel & Friedrich Leisch. 2015. *Miscellaneous functions of the Department of Statistics, Probability Theory Group* (Formerly: E1071), TU Wien. R package version 1.6–7. http://CRAN.R-project.org/package=e1071 (accessed 28 February 2016).

Moore, John M. 2011. *The manuscript tradition of Polybius.* Cambridge: Cambridge University Press.

Mosteller, Frederick, David L. Wallace. 1963. Inference in an authorship problem. *Journal of the American Statistical Association* 58(302). 275–309.

Murtagh, Fionn, and Pierre Legrandre. 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification* 31. 274–295.

Oakes, Michael, P. 2009. Corpus Linguistics and stylometry. In Anke Lüdeling and Merja Kytö (eds.), *Corpus linguistics: an international handbook*, Vol. 2, 1170–1190. Berlin: De Gruyter.

Oakes, Michael, P. 2014. *Natural language processing: literary detective work on the computer.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Pearse, Roger. 2013. *The manuscripts of Polybius*. www.roger-pearse.com/weblog/2013/06/22/the-manuscripts-of-polybius/ (accessed 28 February 2016).

R Core Team. 2015. *R project for statistical computing*. Vienna, Austria. https://www.R-project.org/ (accessed 28 February 2016).

Schindler, Robert M. 1978. The effect of prose context on visual search for letters. *Memory and Cognition* 6(2). 124–130.

Sidorov, Grigori, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh & Liliana Chanona-Hernández. 2012. Syntactic dependency-based N-grams as classification features. In Ildar Batyrshin and Miguel González Mendoza (eds.), *Advances in computational intelligence*, 1–11. Berlin: Springer. http://link.springer.com/chapter/10.1007%2F978-3-642-37798-3_1 (accessed 28 February 2016).

Stamatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology* 60(3). 538–556.

Uzuner, Özlem, Boris Katz. 2005. A comparative study of language models for book and author recognition. In Robert Dale, Kam-Fai Wong, Jian Su, and Oi Yee Kwong (eds.), *International Joint Conference on Natural Language Processing 2005*, 969–980. Berlin: Springer.

van Halteren, Hans, R. Harald Baayen, Fiona Tweedie, Marco Haverkort & Anneke Neijt. 2005. New machine learning methods demonstrate the existence of a human stylome. *Journal of Quantitative Linguistics* 12.1. 65–77.